14th International Conference on Information Fusion
Chicago, Illinois, USA, July 5-8, 2011

Enhanced Target Tracking Through Infrared-Visible Image Fusion

Stephen R. Schnelle
Electrical and Computer Engineering
Rice University
Houston, TX 77005, USA
Stephen.R.Schnelle@rice.edu

Alex Lipchen Chan
Sensors and Electron Devices Directorate
U.S. Army Research Laboratory
Adelphi, MD 20783, USA
Alex.Chan@us.army.mil

Abstract - Video surveillance is an important tool for force protection and law enforcement, and visible and infrared video cameras are the most common imaging sensors used for these purposes. A feasibility study on fusing concurrent visible and infrared imageries using 13 spatial domain and pyramid-based pixel-level fusion algorithms to improve the tracking performance of an existing video surveillance system was performed. Some of the decomposition methods were designed to increase the contrast, whereas some wavelet methods offered shift invariance. The effects of these fusion algorithms on the detection and tracking performance of the given target tracker were examined and compared. Fusion method based on the ratio of low-pass pyramids was shown to offer a superior detection performance at a relatively low computational cost.

Keywords: Image fusion, infrared imagery, visible imagery, target detection, target tracking.

1 Introduction

As sensor technology, network communication, computing power, and digital storage capacity have all dramatically improved, still and video imageries have become the most common and versatile forms of media for capturing, analyzing, and disseminating a variety of information. In many scenarios, useful information is derived from the accurate detection, tracking, and recognition of certain targets of interest in a timely manner. Typical applications of this nature include automatic target recognition systems, force protection surveillance systems, and aerial reconnaissance systems.

Unfortunately, many of these applications involve monitoring adversarial activity in less than ideal environments, which can be particularly challenging to the imaging systems involved. Visible cameras are the prevailing imaging sensors, because they are relatively cheap, easy to use, and capable of producing high quality imagery under favorable conditions. However, visible cameras can be severely affected by many common environmental factors, such as darkness, shadows, fog, cloud, rain, snow, and smoke. Infrared (IR) imaging systems may overcome or alleviate some of these problems, but they are subjected to a number of limitations of their own. IR-specific difficulties include a much lower sensor resolution; drastic diurnal and seasonal changes in target signatures; total loss of non-thermal but important visual features (such as color and text); very low thermal contrast between targets and background under certain combinations of ambient and target temperatures; blockage by visually transparent thermal signal shields (such as car windshields and glass doors)' and much higher costs for purchasing and maintaining the systems. Due to these highly complementary strengths and limitations of visible and IR cameras, more advanced target detection and tracking systems may want to acquire and process both visible and IR imageries concurrently and jointly for critical applications.

Image fusion can be handled at several levels [1]. At the lowest levels, the raw image data can be fused, using either the original signal, or more likely, after the image has been preprocessed, using the resulting pixel values. Pixel-level fusion is very common due to its simplicity and universality, and is the focus of this work as well.

There are many ways to measure the performance of image fusion algorithms, including subjective analysis, complex similarity metrics, signal-to-noise ratio, and tracking performance. Motwani et al. suggested parameters for subjective analysis, but they concluded that subjective measures were not particularly helpful for tracking systems, except in the case of incorporating human feedback into the detection loop [2].

Cvejic et al. discussed a number of objective similarity metrics, including the Piella metric, Petrovic metric, and Bristol metric [3]. The Piella metric measures structured similarity (which is based on luminance, contrast, and structure information) over local window regions and then averages these similarity measures over all windows. Weighting is given to the relative importance of each input image toward the fused image, window by window. The Petrovic metric specifically evaluates edge structure by determining the strength of edge information retained from each of the original images in the fused image. The Bristol metric, in contrast to the Piella metric, uses a slightly different weighting scheme based on the ratio of covariances between the original and fused images.

Cvejic et al. compared the tracking performance of a particle filter based on these objective metrics and found that the tracking performance was actually worsened by

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUL 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Enhanced Target Tracking Through Infrared-Visible Image Fusion	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army Research Laboratory,Sensors and Electron Devices Directorate,Adelphi,MD,20783**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the 14th International Conference on Information Fusion held in Chicago, IL on 5-8 July 2011. Sponsored in part by Office of Naval Research and U.S. Army Research Laboratory.

14. ABSTRACT
Video surveillance is an important tool for force protection and law enforcement, and visible and infrared video cameras are the most common imaging sensors used for these purposes. A feasibility study on fusing concurrent visible and infrared imageries using 13 spatial domain and pyramid-based pixel-level fusion algorithms to improve the tracking performance of an existing video surveillance system was performed. Some of the decomposition methods were designed to increase the contrast, whereas some wavelet methods offered shift invariance. The effects of these fusion algorithms on the detection and tracking performance of the given target tracker were examined and compared. Fusion method based on the ratio of low-pass pyramids was shown to offer a superior detection performance at a relatively low computational cost.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

the fusion of images. Mihaylova et al. of the same research group later adopted a performance metric of normalized overlapping ground truth and tracking system bounding boxes [4]. Their results showed that IR images alone performed just as well or better than most fusion algorithms (including contrast pyramid, dual-tree complex wavelet transform, and discrete wavelet transform) in tracking, while visible spectrum images lagged behind under harsher conditions like occlusions.

There are many possible methods of tracking a moving target, including background subtraction, optical flow, moving energy, and temporal differencing. Because a Force Protection Surveillance System (FPSS) dataset with minimal background interference was available to us, we decided to use the companion FPSS tracker, which is based on background subtraction, to examine the tracking performance of various image fusion methods [5]. Other trackers could be used as well, such as those tracking algorithms surveyed by Trucco and Plakas [6].

In Section 2, the image acquisition and registration process, as well as 13 pixel-level image fusion methods of interest are described. A brief description of the FPSS tracker is provided in Section 3, while the resulting tracking performance of various image fusion methods are presented in Section 4. Finally, some concluding thoughts are given in Section 5.

2 Image Manipulations

2.1 Image Acquisition

To study the effects of fusing visible and IR imagery in detecting and tracking moving targets, we used a large collection of concurrent visible and long-wave infrared (LWIR) video clips called the Second FPSS Dataset [7]. These video clips were collected with the Sentry Personnel Observation Device (SPOD) manufactured by FLIR Systems. As shown in Figure 1, the SPOD includes a LWIR microbolometer and a color visible CCD camera. The LWIR images were acquired with a focal plane array of 320 x 240 pixels in resolution, while the color visible images were captured at the resolution of 460 National Television Standards Committee (NTSC) TV lines.

Figure 1. Sentry Personnel Observation Device.

2.2 Image Registration

Both the original color visible and LWIR images were cropped and scaled to attain a coarse level of co-registration between the corresponding color-LWIR images captured at any given time. The image registration step was necessary, because the color and LWIR cameras of the SPOD were merely bore-sighted into a ruggedized enclosure. They did not share a common optical lens, having slightly different lines of sight, fields-of-view, and image resolutions. Since the distance between these cameras was only a few inches, while the typical ranges to the targets in the FPSS dataset were 50–200 yards, an affine transformation was considered as sufficient.

Image registration can be performed automatically or manually. Although automatic image registration is quite accurate and feasible for images of similar electromagnetic spectrum, registering color and LWIR images is a very difficult task. The effects of automatic and hybrid image registration schemes were explored by Hines et al., but automatic registration was generally not successful [8].

Due to these difficulties, the FPSS dataset was coarsely registered by first manually choosing a large number of salient corresponding markers in many representative pairs of color-LWIR images. The coordinates of these markers were then used to derive the affine transformation between the color and LWIR images through a polynomial fitting process. The maximal usable area could be extracted after applying the affine transformation and avoiding sensor artifacts in both color and LWIR images. The same affine transformation and clipping mechanism were used for the entire dataset. Image patches extracted from the original color and LWIR images were scaled to a common image size of 640 x 480 pixels and stored in JPEG format.

2.3 Image Fusion

2.3.1 Fusion Methods

In this work, we focus on 13 pixel-level image fusion methods, ranging from the simple pixel averaging method to the complicated dual-tree complex wavelet transform, which fall into two broad categories: simple combination and pyramid structure.

The inputs to these image fusion algorithms were those coarsely registered FPSS color and LWIR image pairs, a pair of which is shown in figures 2(a) and 2(b). To allow fusion with LWIR images, the color (RGB) images were first converted to grayscale using a simple weighting of 0.2989R + 0.5870G + 0.1140B, which yielded intensity value but removed hue and saturation information. For many automatic target detection and tracking algorithms, it is indeed more efficient to process grayscale images internally, while providing color outputs for human consumption only. The grayscale visible and LWIR images were manipulated using MATLAB functions to produce various fused images [9, 10].

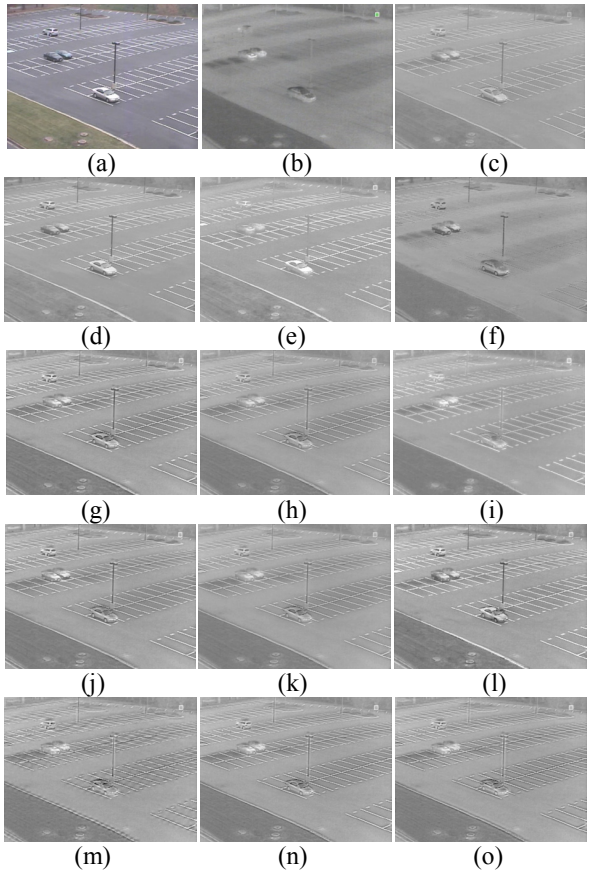

Figure 2. Examples of original (a) Color and (b) LWIR images, as well as fused images produced by (c) Average, (d) PCA-weighted average, (e) Maximum, (f) Minimum, (g) Laplacian, (h) FSD, (i) ROLP, (j) Contrast, (k) Gradient, (l) Morphological, (m) DWT, (n) SIDWT, and (o) DT-CWT fusion methods.

2.3.2 Simple combinations

The most intuitive pixel-level fusion methods examined here are simple averaging, intelligent weighting, and selecting maximum or minimum pixel values. All these methods involve only simple pixel operations, which require traversing the two input images pixel by pixel, leading to simple $O(m \times n)$ operations for an image of size $m \times n$. Pixels $(\mathbf{I}_1)_{ij}$ and $(\mathbf{I}_2)_{ij}$ in images $\mathbf{I}_1$ and $\mathbf{I}_2$ need only be compared against each other once.

In the simple averaging method, fused images were generated by calculating $(\mathbf{I}_f)_{ij} = [(\mathbf{I}_1)_{ij} + (\mathbf{I}_2)_{ij}] / 2$ and an example of the resulting image is shown in Figure 2(c). Because the visible and LWIR images have differing resolutions and salient features, this method tends to muddle the details.

We attempted to boost the influence of the better image by using the Principal Component Analysis (PCA) derived from the covariance matrix of the two input images. We treated each image as a single vector $\mathbf{I}_1$ and $\mathbf{I}_2$, thus creating a 2×2 covariance matrix, when computing the covariance of $[\mathbf{I}_1\ \mathbf{I}_2]$. The normalized eigenvector for the larger eigenvalue provides the necessary weighting: $(\mathbf{I}_f)_{ij} = (\mathbf{v}_k)_1(\mathbf{I}_1)_{ij} + (\mathbf{v}_k)_2(\mathbf{I}_2)_{ij}$, where $\mathbf{v}_k$ represents the eigenvector corresponding to λ_k, the larger one of the two eigenvalues. Generally, the PCA method strongly favors the image with the highest variance, which may or may not contain more informative details. In fact, this selection criterion can be a harmful one when dealing with noisy images. As shown in Figure 2(d), the PCA-weighted fused image closely matches the original visible image due to its higher level of details and variance.

Choosing the maximum pixel value from a pair of LWIR and visible images, $(\mathbf{I}_f)_{ij} = \max[(\mathbf{I}_1)_{ij}, (\mathbf{I}_2)_{ij}]$, may be appropriate to find some hidden targets. A man may be occluded in the visible spectrum, for example, but he can still be located in the LWIR image. For a background subtraction method, it may be desirable to boost the relative intensity of targets through this fusion method, if these targets tend to be brighter than their immediate background. Figure 2(e) shows a fused image generated by the maximum pixel value method.

Choosing the minimum pixel value from a pair of LWIR and visible images, $(\mathbf{I}_f)_{ij} = \min[(\mathbf{I}_1)_{ij}, (\mathbf{I}_2)_{ij}]$, may not be very useful in most cases. As evident from Figure 2(f), this method tends to deemphasize strong foreground objects. In some rare scenarios, this method might be helpful in extracting weak targets (with both weak but detectable visible and LWIR signatures) from a busy background by deemphasizing stronger and brighter background pixels in the neighborhood.

2.3.3 Pyramid Structures

Pyramid decompositions were introduced by Burt and Adelson in 1983 as a compact encoding scheme [11]. The original idea is that a Gaussian kernel (low-pass filter) is applied to the top-level image of a pyramid, $\mathbf{I}_1 * \mathbf{G}_1$, representing the convolution of the image $\mathbf{I}_1$ with a Gaussian blurring matrix $\mathbf{G}_1$. This image is then down-sampled to form the next level of these pyramids. The difference between the low-pass version and its previous-level image represents the high frequency or detail information of the previous-level image. At each step down the pyramid, we continue to filter and down-sample in the same manner. A Laplacian pyramid is formed by computing the difference between each level of the pyramid, iteratively separating an image into low and high frequency components, except that the lowest level contains the remaining low-frequency information.

Since each level is a down-sampled version of the previous level, we need to up-sample and interpolate the decimated version in order to compute the difference between the two adjacent levels. For example, the Laplacian image at level k of $\mathbf{I}_m$, denoted as $(\mathbf{L}_m)_k$, can be computed as $(\mathbf{L}_m)_k = (\mathbf{I}_m)_k - f_{k+1}((\mathbf{I}_m)_{k+1})$, where $f_{k+1}(\ )$ denotes the function consisting of up-sampling and an interpolation filter with similar blurring response as $\mathbf{G}_k$, while k denotes the level of decomposition. As we proceed down the pyramid, $(\mathbf{I}_m)_k$ denotes the blurred and

decimated version of $(\mathbf{I}_m)_{k-1}$. By decomposing each set of the original LWIR and visible images, we form compact representations separated into detail and approximation information. Hence, we can then weight the coefficients in each pyramid. To reconstruct the fused image, we then reverse the decomposition process by combining each level with the successive one. If we select the maximum coefficients between the two pyramids by taking $\max[((\mathbf{L}_1)_k)_{ij}, ((\mathbf{L}_2)_k)_{ij}]$ for each level k and all ij coefficients during this reconstruction process, then a Laplacian fused image is generated (see Figure 2(g)). We could also modify the selection criteria of the algorithm, such as providing more priority to coefficients with larger or more similar coefficients.

Instead of using the maximum coefficients at the lowest level of the pyramid, we may choose to use the LWIR image, the visible image, or a combination of the two at the lowest level as well. If we choose the lowest level LWIR image, this implies that the background for the fused image is built on the LWIR image and detail information from the visible image is only included when these details outweigh those of the LWIR. The Laplacian pyramid is a simple decomposition scheme that assumes very little information about the structure of the image. Implementation details of the Laplacian pyramid include the handling of border effects and ensuring that the image size is a factor of two at each level of decomposition.

A Filter-Subtract-Decimate (FSD) pyramid is similar to the Laplacian pyramid, but the levels are subtracted prior to decimations. Some variations of FSD also make minor adjustments in the synthesis phase. Figure 2(h) shows a fused image produced by the FSD technique proposed by Anderson [12]. Figures 2(g) and 2(h) may look similar, except for a slight shading difference, but their differences in tracking performance could be larger than that.

Ratio-of-low-pass (ROLP) pyramid and contrast pyramid use the ratio of levels of the Gaussian pyramid to produce the next level, rather than the difference [13, 14]. The primary difference between ROLP and contrast pyramids is the use of a local background to normalize the ratio. The contrast pyramid computes a pyramid level as $(\mathbf{L}_m)_k = [(\mathbf{I}_m)_k / f_{k+1}((\mathbf{I}_m)_{k+1})] - 1$, and the offset of 1 is reversed during the reconstruction phase, whereas the ROLP pyramid computes $(\mathbf{L}_m)_k = (\mathbf{I}_m)_k / f_{k+1}((\mathbf{I}_m)_{k-1})$. Note that a small epsilon factor can be added to the denominator to prevent division by zero issues Figures 2(i) and 2(j) shows the resulting fused images from these two methods. These decomposition methods are designed to emphasize the contrast in an image.

The gradient pyramid chooses the largest directional derivative in each of four directions: horizontal, vertical, and the two diagonal directions [15]. These derivatives can be computed using simple matrix operators. Coefficients are selected for each of the four directions independently during the fusion process. An example of fused image produced by the gradient pyramid method is shown in Figure 2(k). These methods preserve orientation information, which can be useful in some applications.

Morphological operations, such as opening and closing, can be applied to the Gaussian pyramid without harmful effects under certain circumstances and result in a morphological pyramid [16]. For example, we can apply the following operations to compute the next set of coefficients from $(\mathbf{I}_m)_k$: morphologically open $(\mathbf{I}_m)_k$ by selecting the smallest nearest neighboring of a pixel, and then the largest in the same region of the resulting image. The resulting image can then be closed by reversing the process, namely, first choosing the largest and then the smallest nearest neighbors. The opening operation will remove small objects, while the closing operation will remove noise and smooth transitions. We decimate the resulting image to obtain our image for the next level of the pyramid, $(\mathbf{I}_m)_{k+1}$. We obtain the pyramid coefficients of level k+1 as the difference between $(\mathbf{I}_m)_k$ and an up-sampled and dilated version of $(\mathbf{I}_m)_{k+1}$. Although the morphological operations may produce interesting results, as shown in Figure 2(l), they are very computationally intensive in nature.

Finally, many specialized pyramid decompositions, such as contourlets and wavelets, separate an image into approximations and detail. We examined a simple discrete wavelet transform (DWT) using the Daubechies Symmetric Spline wavelet, a shift invariant discrete wavelet transform (SIDWT) using the Harr wavelet, as well as a less redundant variant of SIDWT, the dual-tree complex wavelet transform (DT-CWT). Examples of fused images resulting from these three methods are shown in figures 2(m), 2(n), and 2(o), respectively.

The simple DWT can be prone to artifacts as a function of position in the image, which could be particularly problematic when using the FPSS background subtraction tracker to detect motion information. As an object moves slightly, artifacts could shift in the image, resulting in many unnecessary false alarms. Given their shift-invariant property, on the other hand, the SIDWT and DT-CWT are expected to perform better in a tracking task.

3 Target Tracker

Since the FPSS tracker has been developed and adequately tested with the original FPSS dataset, it was chosen for this evaluation work as well.

3.1 Background modeling

The key component of the FPSS tracker is its background modeling and subtraction process, which is depicted in Figure 3. Each input image is first filtered by a stability mask and then channeled through four image buffers of equal size and depth. These buffers are first initialized with the first input image and then gradually updated by other input images in a sequential manner. Each newly arrived set of pixel values replaces the oldest frame in Buffer 1, while the oldest frame of Buffer 1 becomes the newest frame in Buffer 2. The same first-in first-out (FIFO) mechanism of frame-shift and update is applied to all image buffers continuously. Buffers 1 and 3 serve merely as time delays, while the images in Buffers 2

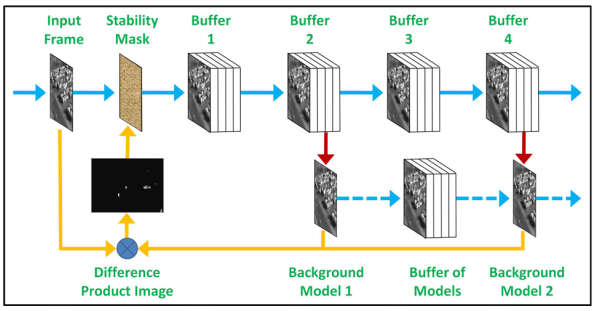

Figure 3. The background modeling and subtraction process in the FPSS tracker.

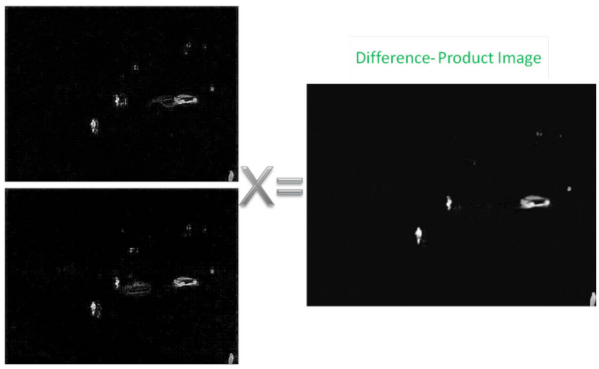

Figure 4. Enhancement of target signatures and suppression of trailing effects and noises via a DPI.

and 4 are used to generate Background Models 1 and 2, respectively. Instead of creating it anew from Buffer 4, Background Model 2 can also be obtained from a buffer of models that is continuously replenished by the outgoing representations of Background Model 1. By subtracting an input frame from these background models, we can obtain two separate difference images.

A difference-product image (DPI) is obtained by multiplying these two difference images pixel by pixel. The product term introduced in this step is useful for the subsequent target detection module, because bright blobs will be generated for all moving targets regardless of the polarity of their original brightness with respect to their immediate background. Although squaring the value of each pixel in the first difference image may achieve the same effects for the first few input frames, the DPI exhibits much better characteristics when the two background models later evolve into two background representations that are clearly disjointed in time.

Typically, each input image frame contains a mostly stable background with a number of small but volatile areas caused by moving objects and other transient events. To prevent rapidly changing foreground pixels from ruining the background model, a stability mask is used to filter out all unstable pixels from the input image frame. Supported by the information provided by the DPIs, this stability mask looks for significant intensity changes based on a predefined threshold of variability and maintains a record of the stability index at each pixel location. Only the stable pixels in a given input image frame are fed to Buffer 1, while those once-stable but now active pixels are blocked and substituted by the corresponding stable pixels available from Buffer 1. Without the stable background models, it will be much harder to detect and extract legitimate moving objects in the scene, and additional false alarms will likely be generated.

This background modeling structure can be extended to include four or more background models for more stable background representations and higher target enhancement capabilities at the expense of additional computational resources. However, an even number of background models should be used to handle issues with target polarity (for example, the LWIR signatures of a human may appear brighter than the background in winter but darker in the summer due to a big change in ambient temperature).

Using multiple disjoint background models to generate a DPI is also a very effective way to remove the problematic "trailing effect" often associated with the background subtraction method. Because the gradually fading trails carved out by the moving objects are now showing up in different parts of the two difference images, they are likely to diminish or disappear when the DPI is formed, as demonstrated in Figure 4. Random noises at different places on the difference images are suppressed in the same way. With the removal or reduction of trailing effect and random noises, the target detector is able to estimate the size and location of the movers much more accurately.

3.2 Target detection and tracking

After a DPI is generated, a morphological operation is used to remove small spikes and fill small gaps in the DPI. Furthermore, a pyramid-means method is used to enhance the centroid and overall silhouette of the moving targets. The moving target detection process begins with finding the brightest pixel on the post-processed DPI, which is usually associated with the most probable moving target in the given input frame. The size of this target is estimated by finding all the surrounding pixels that are deemed connected to the brightest pixel. After a moving target is detected, all the pixels within a rectangular target-sized area surrounding that target are reduced to zero in value, thus excluding them from subsequent detections. This detection mechanism is repeated by finding the next brightest pixel available and it continues until all the pixels are reduced to zeroes, a predefined number of detections are obtained, or other user-defined stopping criteria are met. These stopping criteria may include the minimum size of potential targets, the proportion of overlapping area allowable between adjacent targets, and the "don't care" area, in which all detections should be ignored.

Using the detection results on consecutive input images, tracks of all moving targets are built and maintained. In order to build a meaningful track, a noticeable moving target must appear in multiple contiguous frames in a video sequence. This requirement may not be met when the target is moving across the field of view of the camera at a very short range and/or a very high speed; when the camera is operated at a very low frame rate; when the target is occluded for an extended period of time and/or behind a very large obstacle; or when a combination of these and other detrimental factors occur. The FPSS tracker uses previous locations, velocity, and target size of a moving target to predict the destination of its next movement.

4 Experimental Results

The FPSS dataset consists of 53 short video sequences for a total of 71236 frames depicting various staged suspicious activities around a big parking lot. Ground-truth information (target type and target location) associated with each observable moving target on each image frame was semi-manually generated using a ground-truthing GUI. Using the ground-truth information and the target size estimated by the FPSS tracker, we may compute and compare the tracking performance of the FPSS tracker on the original FPSS dataset and the fused images generated by different image fusion methods.

The ground-truth files associated with a concurrent pair of color-LWIR sequences may vary slightly in their content, because some moving targets may sometimes be observable in one but not both of the imageries. For example, a man walking in a dark area at night can be seen in the LWIR sequence, but is completely obscured in the corresponding color sequence. Since we used the LWIR approximation coefficients during the process of pyramid decompositions and the LWIR ground-truth files usually contain more information on the targets, we chose the LWIR ground-truths files for the purpose of verifying the detections on fused images.

To be qualified for a correct detection or a hit, the ground-truth location must be included in the target-size bounding box estimated by the FPSS tracker for the given detection. Multiple detections on the same target were counted as only one hit, but no additional penalty was imposed in this situation. Multiple detections on a non-target, however, were treated as multiple false alarms (FAs), which would decrease the tracking performance. When multiple targets in proximity were covered by a single detection, it would be treated as multiple hits and would boost the tracking performance. Ground-truth targets that were not included by the bounding box of any detection were regarded as misses that would hurt the tracking performance.

An adjustable acceptance threshold was used to vary the tradeoff between hits and FAs. By gradually lowering the acceptance threshold, the number of hits and FAs increase monotonically. By plotting the FA rate (FAR) (number of incorrect detections per frame) against the hit rate (percentage of true targets that were correctly detected) at different acceptance thresholds, a receiver operating characteristic (ROC) curve is formed. Values closest to the upper-left corner of an ROC curve are the best results, indicating high accuracy with few FAs. To emphasize the critical differences between the ROC curves, we focus on the region of up to 0.1 FA per frame. The ROC curves for the original LWIR and color sequences serve as the benchmark performance metric.

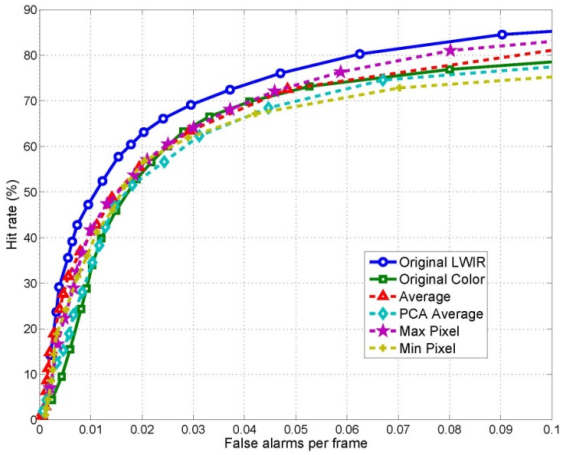

Figure 5. The performance of four simple-combination methods at low FA region.

Figure 5 shows the ROC curves associated with the fused images generated by four simple-combination methods: simple averaging, PCA-weighted averaging, maximum pixel selection, and minimum pixel selection. It is clear that the original LWIR images performed the best with a very low FAR among this group of six candidates. On the other hand, the original color images were lagging behind their LWIR counterparts consistently due to a significant increase in the number of FAs caused by headlight glares and windshield reflections in the evening hours and protracted shadows under the slanted sun. Given the nature of simply averaging or selecting the pixels of the original color and LWIR images by the four simple-combination methods, it is expected that their resulting fused images would perform somewhere between the original color and LWIR images. This is indeed the case for the FAR region of 0.02 or less FAs per frame. As we increase the allowable number of false alarms by lowering the acceptance threshold, the fused images produced by simple averaging and maximum pixel selection methods continue to yield hit rates between the original color and LWIR images, but the fused images generated by PCA-weighted averaging and minimum pixel selection methods gradually fall below the original color images. Hence, there is little to no performance benefit in using images fused with simple combination methods over the original LWIR images. Although it is not shown here, we have found that these trends continue at much higher FARs as well.

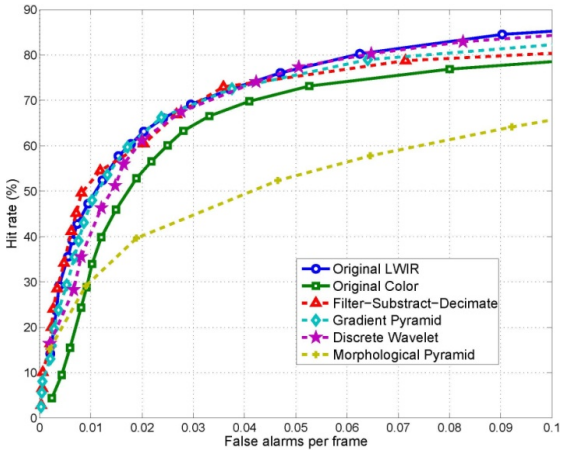

Figure 6. The performance of four inferior pyramid-based fusion methods at low FA region.

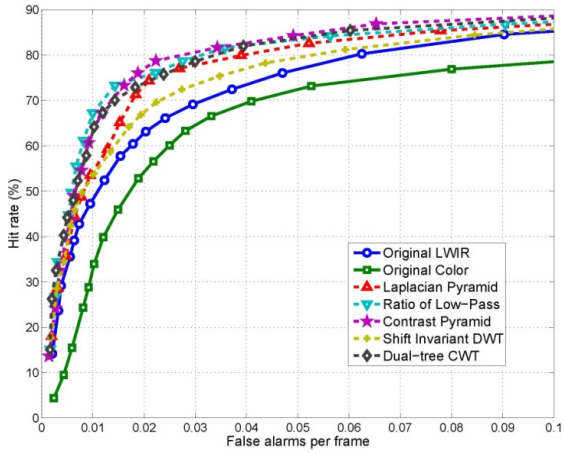

Figure 7. The performance of five superior pyramid-based fusion methods at low FA region.

For the fusion methods based on pyramid structures, they were all performed with an identical set of running parameters, such as using five levels of decomposition and a 7 by 7 neighborhood size when running a saliency/match measure. Based on their resulting ROC curves, these pyramid-based fusion methods can be categorized into two groups for subsequent discussions: four inferior methods and five superior methods. As shown in Table 1, all nine pyramid-based methods are much more computationally intensive than the four simple combination methods, especially the SIDWT, gradient and morphological pyramids. Although the DT-CWT is more than four times more efficient than its more redundant variant, SIDWT, it is still more than twice as slow than the other five simpler pyramid-based methods, three of which are ranked together in the superior pyramid column. More computations do not always generate better results, and as we can see, among the pyramid-based methods there are faster and slower candidates in both the inferior and superior columns of Table 1.

As shown in Figure 6, the FSD, gradient and DWT achieve slightly worse performance than the original LWIR images at low FARs, whereas the morphological pyramid method clearly lags all others under the same conditions. If the plots are extended, we find that the DWT and morphological pyramid methods are able to surpass the LWIR curve at the FAR region of 0.5 FA per frame or higher. Nonetheless, these two methods lack the flexibility to perform in versatile systems.

Finally, there are five pyramid-based fusion methods that can achieve good results on both ends of the ROC curves: the Laplacian, ROLP, contrast, SIDWT, and DT-CWT pyramid methods. As shown in Figure 7, these five fusion methods clearly outperform the original color and LWIR images from the beginning and attain the largest advantage at around 0.02 FA per frame. The advantage of these five fusion methods over the original color and LWIR images is still maintained in the higher FAR region. The SIDWT is slightly behind the other four methods at very high FARs, and since it is much more demanding computationally, it is the least desirable method in this group. Apparently, ROLP works the best in our testing in terms of accuracy and efficiency. Although a 90% hit rate is achievable for almost all these methods with a FAR of 0.5 or less, the marginal benefit in the hit rate quickly decreases in the range of 0.5 to 2 FAs per frame.

5 Conclusion

We explored and exploited the rather complementary natures of two common imaging sensors: LWIR and color visible sensors. Instead of harnessing prior background knowledge and external information sources to perform symbolic level image fusion, we focused on pixel-level image fusion. Therefore, the techniques examined and the results obtained in this work are more readily transferrable to other applications and scenarios that process color and LWIR imageries.

Based on the results generated by the four simple-combination methods examined in this work, we conclude that these methods are not useful, because their performances were worse than using the original LWIR images alone. Among the 9 pyramid-based fusion methods, the gradient and FSD methods are even worse than the simple-combination methods, because they required 10–60 times more computational resources but performed worse at high FAR region. The morphological and DWT methods are slightly better than the gradient and FSD methods, primarily because they managed to outperform LWIR in the high FAR region. On the other hand, the Laplacian, ROLP, contrast, SIDWT, and DT-CWT are considered superior fusion methods, because they consistently outperformed LWIR in every FAR region, though the SIDWT required substantially more computational resources. ROLP and contrast methods can be considered the best with the FPSS tracker, as their ROC curves are consistently superior while their computational needs are among the lowest of the pyramid-based methods.

Table 1. CPU time (seconds) spent on making 30 fused images using MATLAB code on a Dell T7400 workstation.

Simple combination	CPU time	Inferior pyramid	CPU time	Superior pyramid	CPU time
Simple average	1.280	FSD	21.670	Laplacian	24.040
PCA average	2.030	Gradient	78.970	ROLP	23.050
Max pixel	1.560	DWT	22.740	Contrast	23.240
Min pixel	1.840	Morpho	62.530	SIDWT	209.600
				DT-CWT	49.940

One simple possibility for improving performance may be to treat each color image as three separate images (RGB) and fuse the set of four images together. The fusion algorithms do not limit the number of images that can be fused. Short-wave infrared (SWIR) and hyperspectral imageries could also be included if they are properly co-registered. Performance may also be further improved by linking the fusion process with the tracking algorithm, through which the information that is critical to the tracker may be better preserved or enhanced. For instance, a region-based segmentation algorithm may be incorporated into a DT-CWT fusion process [17], exploiting the limited redundancy in DT-CWT and tying the feature level and pixel-level fusion algorithms together.

Acknowledgements

This work was partially supported by the ARO MURI W311NF-07-1-0185 grant, National Science Foundation Graduate Fellowship Program, National Defense Science and Engineering Graduate Fellowship Program, and the Texas Instruments Leadership University Program.

References

[1] M. Smith and J. Heather, "Review of Image Fusion Technology in 2005," *Proc. SPIE Thermosense*, vol. 5782, pp. 29–45, Mar 2005.

[2] M. Motwani, N. Tirpankar, R. Motwani, M. Nicolescu, and F. Harris, "Towards Benchmarking of Video Motion Tracking Algorithms," *Int. Conf. Signal Acquisition and Processing*, pp. 215-219, 2010.

[3] N. Cvejic, S. G. Nikolov, H. D. Knowles, A. Loza, A. Achim, D. R. Bull, and C. N. Canagarajah, "The Effect of Pixel-Level Fusion on Object Tracking in Multi-Sensor Surveillance Video," *IEEE Conf. Computer Vision and Pattern Recognition*, vol. 372, pp. 1-7, 2007.

[4] L. Mihaylova, A. Loza, S. Nikolov, and J. J. Lewis, "The Influence of Multi-Sensor Video Fusion on Object Tracking Using a Particle Filter," *Proc. 2nd Workshop on Multiple Sensor Data Fusion*, pp. 354-358, 2006.

[5] A. L. Chan, "A Robust Target Tracking Algorithm for FLIR Imagery," *Proc. SPIE Automatic Target Recognition*, vol. 7696, pp. 1-11, May 2010.

[6] E. Trucco and K. Plakas, "Video Tracking: A Concise Survey," *IEEE Journal of Oceanic Engineering*, vol. 31, pp. 520-529, 2006.

[7] A. L. Chan, "A Description on the Second Dataset of the U.S. Army Research Laboratory Force Protection Surveillance System," *Memo Report ARL-MR-0670*, U.S. Army Research Laboratory, pp. 1–28, 2007.

[8] G. Hines, Z. Rahman, D. Jobson, and G. Woodell, "Multi-image Registration for an Enhanced Vision System," *Proc SPIE Visual Information Processing*, vol. 5108, pp. 231-241, Aug 2003.

[9] O. Rockinger, "MATLAB Image Fusion Toolbox," http://www.metapix.de/indexp.htm,1999 (accessed 2010).

[10] S. Cai and K. Li, "Matlab Implementation of Wavelet Transforms," http://taco.poly.edu/WaveletSoftware/ references.html (accessed 2010).

[11] P. J. Burt and E. H. Adelson, "The Laplacian Pyramid as a Compact Image Code," *IEEE Trans. Communications*, vol. 31, pp. 532-540, 1983.

[12] H. Anderson, "A filter-subtract-decimate hierarchical pyramid signal analyzing and synthesizing technique," U.S. Patent 718 104, 1987.

[13] A. Toet, "Image fusion by a ratio of low-pass pyramid," *Pattern Recognition Letters*, vol. 9, pp. 245-253, 1996.

[14] A. Toet, J. J. van Ruyven, and J. M. Valeton, "Merging Thermal and Visual Images by a Contrast Pyramid," *Optical Engineering*, 28 (7), 789-792, 1989.

[15] P. Burt, "A gradient pyramid basis for pattern selective image fusion," Society for Information Displays (SID) Int. Symp. *Digest of Technical Papers*, vol. 23, pp. 467-470, 1992.

[16] L. C. Ramac, M. K. Uner, and P. K. Varshney, "Morphological filters and wavelet based image fusion for concealed weapon detection," *Proc. SPIE Sensor Fusion*, vol. 3376, pp. 110-119, 1998.

[17] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel- and Region-Based Image Fusion with Complex Wavelets," *Information Fusion*, vol. 8, pp. 119-130, 2007.